Issues Associated with the Design of a National Probability Sample for Human Exposure Assessment

Trena M. Ezzati-Rice and Robert S. Murphy

National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland

Data obtained from national probability sample surveys provide important information on the prevalence of various health conditions and distributions of physical and biochemical characteristics of the U.S. population. The sample design of a survey specifies how sampling from a designated population over a stated period is to be accomplished. A survey's analytical objectives and interests—in particular subpopulations—affect the sample design strategy. Selected subdomains of the population often must be oversampled so that estimates can be made with acceptable precision. This article addresses sample design considerations for a national probability sample for human tissue monitoring and specimen banking. Among the sampling issues addressed are the oversampling of special populations e.g., minority groups and at-risk groups such as low income or elderly persons; geographic coverage; and sample size considerations. The sample design for a major health survey, the Third National Health and Nutrition Examination Survey (NHANES III), is used to illustrate a complex, multistage probability sample design and to highlight some of the sampling issues discussed in this article. — Environ Health Perspect 103(Suppl 3):55–60 (1995)

Key words: survey design, oversampling, stratification, multistage sampling

Introduction

National probability sample surveys provide important information on the prevalence of various health conditions and distributions of physical and biochemical characteristics of the United States population as well as providing data on the relationship between risk factors and selected conditions (*1,2*). Recent advances in industrial technology have resulted in the release of a diverse number of toxic substances into the environment resulting in an increasing interest in monitoring and assessing the exposure of the human population to environmental pollutants. The collection of human tissue samples as part of national sample health surveys could provide a unique source of data to help establish public health priorities in environmental health and to help define guidelines for the prevention of environment-related diseases. Further, tissue specimen banks could be useful for future epidemiologic and risk assessment studies and for evaluation of public health prevention efforts.

This paper was presented at the Conference on Human Tissue Monitoring and Specimen Banking: Opportunities for Exposure Assessment, Risk Assessment, and Epidemiologic Research held 30 March – 1 April 1993 in Research Triangle Park, North Carolina.

Address correspondence to Trena M. Ezzati-Rice, National Center for Health Statistics, Centers for Disease Control and Prevention, 6525 Belcrest Road, Hyattsville, MD 20782. Telephone (301) 436-7022. Fax (301) 436-7955.

The design and conduct of current and previous national probability health surveys could help guide the development of national surveys for human exposure assessment. This article addresses sample design considerations for a national probability sample for human tissue monitoring and specimen banking. Among the sampling issues addressed are defining the sampling frame and sampling units; stratification; sampling of special populations (e.g., minority groups and at-risk groups); geographic coverage; and factors affecting sample size and its allocation over regions, subdomains, and strata. The Third National Health and Nutrition Examination Survey (NHANES III) is used to illustrate a complex, multistage sample design for a major health survey and to highlight some of the sampling considerations discussed in this article.

Survey Design Issues

The planning of national probability sample surveys requires survey designers to address a number of complex issues. Sample design is only one of several interrelated aspects involved in designing a sample survey. In an attempt to achieve an efficient and economic sample design, the sample designer must have knowledge of the objectives to be served by the survey's results. Thus, the development of a well-defined set of goals and objectives is the first, and perhaps, most critical, task of planning a national sample survey for human exposure assessment because it determines the alternatives for the sampling units and frame, the actual sample design and selection procedures, and estimation methods.

Other important issues that must be addressed when designing a sample survey include defining the target population about which estimates are to be made, including any subdomains of the population and the level of precision desired for the estimates (total and subdomain estimates). In addition, the geographic coverage for the survey must be determined. A number of operational issues also must be considered in the design of a survey, including the types of data to be collected, the methods for obtaining the needed survey data, the design and contents of questionnaires, and the time frame in which the survey is to be conducted. Further, for environmental risk assessment, issues surrounding the long-term storage and management of biologic specimens must be addressed. Finally, sources of potential nonsampling error (e.g., nonresponse and noncoverage) should also be anticipated and efforts made to minimize their impact on survey estimates. After consideration of these issues, along with the resources available to conduct the survey, the sample design for the survey can be developed.

Sample Design Issues

The sample design for a survey includes the sampling plan; the set of procedures by which the sample is selected from the target population; the estimation procedures, the

set of algorithms for estimating population values from the sample and estimating the reliability of the estimates (*3,4*). Development of the sample design for a survey involves not only developing the sampling frame, but determining the total sample size and its allocation by areas of the country (e.g., region or state); subdomains (e.g, the elderly, children, Mexican Americans), and strata; and any clustering of the sample to reduce survey costs. A good sample design produces estimates that are unbiased and reliable. It also includes specification for feasible survey operations.

The strategy for selection of an appropriate sample selection method depends first on the survey's primary objectives. Other important considerations, however, include the degree of precision required for estimates and the sample size needed to meet the precision requirements. The sample size for the survey directly affects the cost of the survey; therefore, the sample size should be selected so as to maximize the reliability (accuracy) of the survey results and at the same time to minimize the cost of the survey. Other practical issues are the staff resources for the field operations, the time frame for the survey, and the total budget.

In the absence of any convenient ready-made sample frame, a multistage area sample is generally chosen for sampling people at large or for populations whose members are widely scattered. A multistage area sample is typically used for large-scale government-sponsored household surveys like the National Health Interview Survey and the Third National Health and Nutrition Examination Survey (*1,2*). This type of sampling provides a great deal of flexibility in the kind, number, and size of the sampling units at each stage of selection as well as the number of stages to use.

A national probability sample design typically involves stratification, oversampling, and differential probabilities of selection. Stratification is usually undertaken to improve sampling efficiency, but it also offers some other benefits (*5*). In addition to improving the accuracy of survey estimates, stratification can simplify survey planning and administration. Further, it can help to form strata for which separate estimates are needed. For example, in a national survey of human exposure assessment, separate estimates by region may be desired; therefore, regions could be treated as strata with sample selection from each so as to permit efficient comparisons. Another distinct advantage of a multistage design is that it provides the opportunity for stratification at various stages of the design for desired subdomains of the population. Differences in the health status of racial and ethnic minorities is a key public health concern. Although not excluded from the target population, small numbers of selected subpopulations — for example, blacks and Mexican Americans — are included in population surveys based on a probability proportional to size-sample design. Oversampling specific population groups is often used to address this limitation. More specifically, geographic stratification is one method often used to sample racial or ethnic minorities that are clustered in the population to reduce survey costs and for administrative efficiency. However, for many small (rare) population subgroups (e.g., Native Americans or persons 65 years and older), there are methodologic barriers to employing oversampling strategies. These include geographic diffusion and the fact that screening for selected subgroups is often costly. Some alternative methods for sampling special populations (small subgroups of the population of analytical interest) are addressed in the next section.

Sampling of Special Populations

One of the goals of national surveillance for human exposure would likely be to assess the prevalence of exposure and levels of exposure of the United States population to carcinogens and other toxic substances. Another possible objective might be to identify population subgroups at increased risk of exposure to carcinogens and other toxic substances (e.g., minorities, women, children, elderly, or persons of lower socioeconomic status). Therefore, in the early planning stage of a survey, specific subgroups of interest should be identified and prioritized or given equal interest.

There are various sample selection techniques to ensure that various subpopulations are adequately represented in a selected sample. Most of the widely used methods give different probabilities of selection to various subdomains. This results in a sample in which the proportions of particular units vary from their occurrence in the population sampled. However, other alternative methods deserve mention.

If a complete list of a special population exists, then it should be used and no other sampling technique would be necessary. However, this is rarely the case and when lists do exist, coverage is often incomplete or unknown. Sometimes the use of multiple partial lists may provide some coverage for subpopulations of interest. However, some screening would likely still be necessary. In some instances, the use of previously collected data from another sample survey can be used to provide a sample of a special population (*6*). This linkage of surveys might be considered if the research called for including a sample of persons with a selected health condition. For example, information on chronic health conditions collected as part of the National Health Interview Survey (NHIS) might provide a partial list of persons with a selected target condition of interest. Another use of this strategy might be applied if a previously conducted national health survey had determined ethnicity for each sample person. A sample of, say, Hispanics could then be selected from that sample for a new study. However, there are some limitations to this approach, such as the quality of the data, tracing of people who have moved, and privacy concerns of respondents (*7*). Multiplicity or network sampling, obtaining information from respondents about persons with whom they are connected, can be effective for some rare populations or conditions; but it is not appropriate in all cases (*8*).

If special populations are clustered, there are several alternatives for sampling that can reduce the survey costs. If geographic segments with no members of a special population are known in advance, for example from census data, screening in these segments could be eliminated, resulting in a cost savings. However, for some special populations census data may not be available, in which case telephone or mail screening could be used to help determine the nonzero segments. Other methods discussed include face-to-face screening and the use of lists to identify nonzero geographic segments (*7*).

Many national surveys have an interest in producing separate estimates for selected subgroups of the population (e.g., minorities, children, elderly, women, and persons below the poverty level). Producing statistically reliable estimates for small subdomains of the population often requires that these subgroups of the population be oversampled. Oversampling blacks and especially Hispanics poses a special challenge. An effective method for oversampling subdomains of the population is to stratify geographic areas by concentration of the minority population and to oversample those areas with high concentrations. This design feature has been used in the Third National Health and Nutrition Examination Survey (NHANES III) and the National Survey of Family Growth, Cycles II and III (*1,9,10*). The advantage of this procedure is that it increases the reliability

of statistics for the minorities and in most cases will have a modest effect on estimates for the total population. The disadvantage of this option is the extensive household screening still required if minority statistics are desired for specific age domains that are particularly rare. Research has shown that screening for sufficient numbers of black and Hispanic Americans who are 65 years and older is very costly. Therefore, another method to reduce screening is to obtain lists that contain a high proportion of the special population domain. The list-frame sample can then be used to supplement the area sample (dual-frame sampling). Potential list-frame sources for the elderly include Medicare files maintained by the Health Care Financing Administration and administrative files maintained by the Social Security Administration (SSA). One complication in the use of these list files for oversampling elderly Hispanics is that no specific Hispanic identifier exists; therefore, Hispanic surname would need to be used to identify elderly Hispanics from the list frame. A couple of other issues must be considered before implementing a multiple frame sampling approach. First, survey costs and complexity increase with a multiple frame survey, and issues of coverage emerge. Second, complex estimation techniques are required for appropriate statistical analyses. More specifically, the probability of selecting a sample person from each frame used must be determined and any duplication of persons on multiple frames must be accounted for. Therefore, matching of frames becomes an important operation. A study to investigate the use of SSA files to oversample elderly minorities in the National Health Interview Survey has been previously described (*11*).

In many health surveys there is an interest in specific at-risk populations such as persons whose income is below the poverty level. This might be the case for environmental exposure as well. Therefore, oversampling geographic areas by income class might be a consideration. However, research has shown that poverty is not sufficiently concentrated for stratification and the oversampling of high-density poverty areas to make it cost effective (*12*). The primary reason is that low-income persons do not live in sufficiently high areas of concentration. Furthermore, since the only source of data for stratification by income is the most recent decennial census, these data become outdated for intercensal years, especially towards the end of a decade. A further complicating factor is the poor quality of income reporting in screening interviews. Also, any oversampling procedure that has to rely on screening to identify low-income persons, by definition would need to be short and would then be subject to measurement error.

Previous research on sampling rare populations indicates that very high concentrations of the rare population must exist for differential sampling rates among strata to be effective (*13*). Further, it has been shown that income class is an example of a subpopulation for which disproportionate sampling is not particularly effective (*3,14*).

To illustrate some of the sample design issues discussed in the previous sections of this article, the design of a major national health survey will be described in the next section. Although no one study can serve as the protocol for another since each has separate objectives, precision requirements, and budgetary constraints, some features of previous national probability sample surveys might be applicable to the design of a national survey for human exposure assessment. The design of the Third National Health and Nutrition Examination Survey (NHANES III) is therefore useful in illustrating a multistage area probability sample design as well as the oversampling of selected subgroups of the population.

Example of a National Random Sample Design

General Description of NHANES III

The National Health and Nutrition Examination Survey (NHANES) is a periodic survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHANES collects important nutritional and health-related data on the civilian, noninstitutionalized United States population and important subgroups. The Third National Health and Nutrition Examination Survey (NHANES III) is the seventh in a series of similar surveys conducted by NCHS since the 1960s (*15–22*). NHANES III is designed to provide national statistics on health and nutritional status for the civilian noninstitutionalized population. Sociodemographic and medical history data are obtained through personal household interviews, while physical measurements, physiologic tests, and biochemical measurements are collected through standardized physical examinations in specially equipped mobile examination centers that are transported to each survey location. NHANES III was partitioned into two 3-year surveys —Phase 1 (1988–1991) and Phase 2 (1991–1994)— to provide national estimates for each 3-year period as well as for all 6 years.

The broad sample design specifications for NHANES III called for statistically reliable estimates for 52 detailed age, sex, and race/ethnicity subdomains (Table 1). The level of precision required for each subdomain was a relative standard error of 30% or less for a 10% statistic. In addition, we wanted to be able to detect differences of 10% with a Type I error of 0.05 or less and a Type II error of 0.10 or less. A description of the target population, expected sample sizes, stages of the hierarchical design, and estimation procedures are described briefly below.

Target Population

The NHANES III sample was designed to cover the noninstitutionalized population of the United States aged 2 months and older. Children under 5 years of age, adults aged 60 years and older, and both black and Mexican American persons are sampled at higher rates than other persons to provide reliable estimates for these important demographic subpopulations.

Sample Size

The sample size was fixed with regard to the resources available and the time period for the conduct of the survey. The sample size was determined in part from past NHANES experience taking into account patterns of nonresponse and the time required to conduct the examination portion of the survey. The desired sample size was 40,000 sample persons with 12,000 blacks, 12,000 Mexican Americans, and 16,000 whites and all others. Table 2 shows the sample sizes, by race/ethnicity, that are expected to be

Table 1. Analytic subdomains classified by race/ethnicity and age.

Black	White and all other	Mexican–American[a]
2–35 months	2–11 months 12–35 months	2–35 months
3–5 years	3–5 years	3–5 years
6–11 years	6–11 years	6–11 years
12–19 years	12–19 years	12–19 years
20–39 years	20–29 years 30–39 years	20–39 years
40–59 years	40–49 years 50–59 years	40–59 years
60 years and over	60–69 years	60 years and over
	70–79 years	
	80 years and over	

[a]Mexican–Americans can be any race. Note: The analytic subdomains are for males and females separately. Source: Third National Health and Nutrition Examination Survey, 1988 to 1994.

Table 2. Expected sample sizes with and without oversampling.

Race/ethnicity	With oversampling		Without oversampling	
	Number	Percent	Number	Percent
Total	40,000	100	40,000	100
Black	12,000	30	4,800	12
Mexican–American	12,000	30	2,400	6
White and all others	16,000	40	32,800	82

Source: Third National Health and Nutrition Examination Survey, 1988 to 1994.

achieved with geographic stratification and oversampling and the corresponding sample sizes that would have been obtained without any oversampling. Assuming a 75% response rate, the sample is expected to yield a total of 30,000 examined persons.

Type of Design

Like most national random samples, NHANES III is a multistage stratified design. The four stages of the design are shown in Table 3.

One of the operational advantages of a multistage national probability survey design is that a relatively small number of areas can be designated for the conduct of the survey, which limits the number of sample areas to which survey personnel must travel. For example, in NHANES III the entire United States (including Alaska and Hawaii) was divided into approximately 2812 geographic areas or primary sampling units (PSUs), most of which consisted of individual counties. After the selection of 13 very large PSUs with certainty, the remaining PSUs were grouped into 34 strata according to region, SMSA status, race/ethnicity, and income and 2 PSUs were selected per strata. These 81 PSUs were the first-stage units of selection. The noncertainty PSUs were selected to take into account the need for reliable statistics for black and Mexican–American persons. In each of the sample PSUs, successive stages of sample selection included segments (census enumeration districts or block groups), households, and sample individuals. To reduce the cost of screening necessary to locate the desired Mexican Americans for the sample, area segments consisting of census block groups and enumeration districts are stratified by the percent of the population that is Mexican American, with a higher rate of selection used in strata containing 3% or greater Mexican American population. Households are also sampled at variable rates depending on the concentration of Mexican Americans within the stratum. Within households, children under 5 years, persons over 60, blacks, and Mexican Americans are oversampled. A detailed description of the NHANES III sample design has been previously published (*1*).

Estimation Procedures

The NHANES, like most sample surveys, experiences unit or total nonresponse despite special procedures designed to maximize response rates. For NHANES III, these procedures include extensive publicity in each survey location, a home examination especially targeted for the older population, a remuneration to all examined participants, and a report of major medical findings. Since NHANES includes both an interview and an examination component, two levels of unit nonresponse occur. That is, some persons randomly selected for the survey refuse to be interviewed and some who are interviewed refuse the examination portion of the survey. NHANES III-phase 1, conducted from 1988 to 1991, included 20,277 sample persons. In-person household interviews were conducted with 17,464 persons (86%) and physical examinations were conducted with 15,864 persons (78%). Table 4 shows the examination response rates for males and females by age and race/ethnicity. The examination response rate was highest for the two minority subgroups, and response rates decreased with increasing age for both males and females.

Two features of the NHANES III design must be taken into account in any analysis of the data collected. The first is the use of sample weights so that correct national population estimates can be produced. For NHANES III, the final analysis weights incorporate the selection probabilities and include adjustments for nonresponse. The nonresponse adjusted weights are further poststratified by age, gender, and race/ethnicity to account for noncoverage and to bring the final national estimates in line with known population counts. The weighting procedure for NHANES III has been previously described (*23,24*). The second feature of the design that must be taken into account is the strata and primary sampling units from the complex sample design to estimate variances and test for statistical significance.

Discussion

Currently, no national data are available on the prevalence of exposure of the United States population to various toxic substances. A limited number of measurements for toxic substances have been done as part of the NHANES including lead, cotinine, selected pesticides, cadmium, benzene, styrene, and a few others. Clearly, national probability surveys to determine the exposure to various toxic substances could help identify at-risk populations and establish prevention programs. Furthermore, biologic specimen banking could provide a valuable resource to permit future laboratory analyses for the prevalence of toxic substances of emerging importance with respect to disease risk and

Table 3. Four stages of multistage stratified design.

Stage	Sampling unit	Stratification
1	Counties	Region, SMSA status, race/ethnicity, income
2	Segments	Mexican–American density strata
3	Households	Minority concentration
4	Persons	Age, gender, race/ethnicity

Source: Third National Health and National Health and Nutrition Examination Survey, 1988 to 1994.

Table 4. Examination response rate by gender, age, race/ethnicity.

Age	Males, percentage				Females, percentage			
	All	Blacks	Mex–Am	W/other	All	Blacks	Mex–Am	W/other
≤5	88	91	87	86	88	92	89	86
6–19	84	85	85	81	85	88	88	80
20–44	74	79	76	68	81	86	80	77
45–59	72	73	73	69	74	78	75	72
60–74	71	73	72	70	67	67	69	65
75+	67	76	63	66	62	71	68	59
All ages	78	82	80	74	79	84	83	74
Total age and gender	78	83	81	74				

Mex–Am, Mexican–Americans; W, white. Source: Third National Health and Nutrition Examination Survey-Phase 1, 1988 to 1991.

for determining the prevalence of environmental agents as new techniques are developed to measure environmental exposure. This article has addressed a number of sample design issues that need to be considered in the design of possible surveys of human exposure assessment to ensure that the survey goals and objectives can be met. Surveys designed for the general population likely cannot address all of the risks of exposure for numerous at-risk groups. Therefore, in addition to national probability sample surveys to provide data on environmental exposures for the United States population, consideration should also be given to conducting selected special studies of potentially exposed populations, such as persons living in certain agricultural areas or industrial areas.

REFERENCES

1. Ezzati TM, Massey JT, Waksberg J, Chu A, Maurer K. Sample Design: Third National Health and Nutrition Examination Survey. National Center for Health Statistics. Vital Health Stat 2(113), 1992.
2. Massey JT, Moore TF, Parsons VL, Tadros W. Design and Estimation for the National Health Interview Survey. National Center for Health Statistics. Vital Health Stat 2(110), 1989.
3. Kish L. Survey Sampling. New York:John Wiley & Sons, 1965.
4. Levy PS, Lemeshow S. Sampling for Health Professionals. Belmont, CA:Lifetime Learning Publications, 1980.
5. Foreman EK. Survey Sampling Principles. New York:Marcel Dekker, 1991.
6. Sudman S. Reducing the Cost of Surveys. Chicago:Aldine, 1967.
7. Sudman S, Kalton G. New developments in the sampling of special population. Ann Rev Sociol 12:401–429 (1986).
8. Sirken MG. Household surveys with multiplicity. J Am Stat Assoc 65:257–266 (1970).
9. Grady WR. National Survey of Family Growth, Cycle 2: Sample Design, Estimation Procedures, and Variance Estimation. National Center for Health Statistics. Vital Health Stat 2(87), 1981.
10. Bachrach CA, Horn MC, Mosher WD, and Shimizu I. National Survey of Family Growth, Cycle III: Sample design, Weighting, and Variance Estimation. National Center for Health Statistics. Vital Health Stat 2(98), 1985.
11. Ezzati TM, Hoffman K, Judkins DR, Massey JT, Moore TF. A dual frame design for sampling elderly minorities and persons with disabilities. In: Statistics in Medicine, Vol 13.
12. Chu A, Lannom L, Morganstein D, and Waksberg J. NHANES III Methods Research Final Report. Contract No 282-86-0042. Westat, Inc, Rockville, MD, 1989.
13. Kalton G, Anderson DW. Sampling rare populations. J Roy Stat Soc, Ser A 149:65–82 (1986).
14. Waksberg J. The effect of stratification with differential sampling rates on attributes of subsets of the population. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1973;429–434.
15. National Center for Health Statistics. Plan and Initial Program of the Health Examination Survey. National Center for Health Statistics. Vital Health Stat 1(4), 1965.
16. National Center for Health Statistics. Plan, Operation, and Response Results of a Program of Children's Examinations. National Center for Health Statistics. Vital Health Stat 1(5), 1968.
17. National Center for Health Statistics. Plan and Operation of a Health Examination Survey of U.S. Youths 12–17 Years of Age. National Center for Health Statistics. Vital Health Stat 1(8), 1969.
18. Miller HW. Plan and Operation of the Health and Nutrition Examination Survey, United States, 1971–73. National Center for Health Statistics. Vital Health Stat 1(10a), 1978.
19. National Center for Health Statistics. Plan and Operation of the Health and Nutrition Examination Survey, United States, 1971–73. National Center for Health Statistics. Vital Health Stat 1(10b), 1977.
20. Engel A, Murphy RS, Maurer K, Collins E. Plan and operation of the HANES I Augmentation Survey of Adults 25–74 Years, United States, 1974–75. National Center for Health Statistics. Vital Health Stat 1(14), 1978.
21. McDowell A, Engel A, Massey JT, Maurer K. Plan and Operation of the second National Health and Nutrition Examination Survey, 1976-80. National Center for Health Statistics. Vital Health Stat 1(15), 1981.
22. Maurer KR. Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. National Center for Health Statistics. Vital Health Stat 1(19), 1985.
23. Ezzati TM, Khare M. Consideration of health variables to adjust sampling weights for nonresponse in a national health survey. Proceedings of the Social Statistics Section of the American Statistical Association, 1991;202–208,.
24. Ezzati TM, Khare M. Nonresponse adjustments in a national health survey. In: Proceedings of the Survey Research Methods Section of the American Statistical Association, 1992;339–344.